Likelihood based twin fraction estimation

P.H. Zwart,[a]* R.J. Read,[b] R.W. Grosse-Kunstleve[a] and P.D. Adams[a]

[a]*Lawrence Berkeley National Laboratories, One Cyclotron road, Berkeley, CA 94720, USA, and* [b]*University of Cambridge, Department of Hematology, Cambridge UK.*

E-mail: PHZwart@lbl.gov

(Received 0 XXXXXXX 0000; accepted 0 XXXXXXX 0000)

twinning; likelihood

Abstract

Likelihood based techniques for determining the twin fraction are described.

1. Introduction

Twinning is relatively common phenomena in protein crystallography (XXXX). Twinning can occur when the symmetry of the crystal lattice has a higher symmetry than the symmetry of the (untwinned) intensities. If the (approximate) symmetry of the lattice is a super group of the point group of the intensities, diffraction patterns of differently oriented crystal domains can overlap (almost) perfectly (XXXX). As the intensities from a twinned crystal can be treated as the sum of the intensities of two twin related miller indices (XXXX), the presence of twinning can often be detected by the departure of the observed intensity statistics from what is known from Wilson statistics (XXXX, XXXX, XXXX), or from standards derived from a database analysis (Zwart *et al.*, 2006). Further introduction into twinning and twinning related nomenclature can be found in a number of excellent introductions in twinning (XXXX, XXXX, XXXX). as the effect of twinning on the intensities can be described by a summation of the intensities of the individual twin domains, we can write

$$\begin{pmatrix} J_1 \\ J_2 \end{pmatrix} = \begin{pmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{pmatrix} \begin{pmatrix} I_1 \\ I_2 \end{pmatrix} \tag{1}$$

J_1, J_2 denote intensities of twinned data, and I_1, I_2 denote intensities of the untwinned data. The effect of α on the distribution of the intensity J is shown in figure 1. Note that both I and J are quantities free of experimental errors.

As can be seen from figure 1, the presence of twinning can be detected form the moments of the distribution of structure factor amplitudes. The moments of the twinnned intensities J depend on the twin fraction in the following manner (appendix A):

$$\mathbb{E}[J^k] = \frac{(\epsilon\sigma_p(1-\alpha))^{1+k} - (\epsilon\sigma_p\alpha)^{1+k}}{(1-2\alpha)\epsilon\sigma_p}\Gamma[1+k] \tag{2}$$

From expression (2), it can be clearly seen that

$$\frac{\mathbb{E}[J^2]}{\mathbb{E}[J]^2} = 2(1-\alpha(1-\alpha)) \tag{3}$$

and results in the familiar Wilson intensity ratios of 2 and 1.5 for untwinned ($\alpha = 0$) and perfectly twinned acentric data ($\alpha = 0.5$).

2. Background

Introduce the distributions and expressions as derived in the appendices. We will use them in the upcomming sections.

3. Classic methods of twin fraction estimation

3.1. Murray-Rust plot

Not commonly used, but instructive. same as britton plot, less numerically accesable.

3.2. Britton plot

The britton plot is a relative straightforward method that can be used to determine the twin fraction - Show integral and why we end up with a straight line. - show breakdown with presence of NCS - talk about absense of use of experimental errors

3.3. H-test

As above, but on top of that, do numerical simulation for various twin fractions and values of D_{ncs}.

4. Results

A database analyses of various twin fraction nestimation methods. - compare britton, H and standard ML estimates (from xtriage run on PDB subset) - compare twin fraction as estimated when Dncs is included. Show that twin fraction comes out lower than what is obtained from other methods (including model refinement), but that Dncs is a good estimator for the correlation between error free twin related amplitudes and can be possibly used as an easy way of detection NCS parallel to the twin axis.

5. Discussion and conclusions

etc etc

Appendix A
Independent twin related amplitudes

A.1. Error free quantities

When twin related intensities I_1 and I_2 are assumed to be independent, one can write

$$f(I_1, I_2) = \frac{1}{(\epsilon\sigma_p)^2} \exp\left[-\frac{I_1 + I_2}{\epsilon\sigma_p}\right] \tag{4}$$

Twinning can be introduced by using the inverse of expression (1) and introduction of the appropriate Jacobian (XXXX):

$$f(J_1, J_2) = \frac{1}{(\epsilon\sigma_p)^2(1-2\alpha)} \exp\left[-\frac{J_1 + J_2}{\epsilon\sigma_p}\right] \tag{5}$$

Note that the transformation (twinning) results in a new domain on which the density function is defined. The domain for the untwinned data is defined by

$$I_1 \geq 0 \tag{6}$$

$$I_2 \geq 0 \tag{7}$$

Upon application of the inverse of expression (1), one obtains

$$\frac{\alpha}{1-\alpha} J_1 + \frac{1-\alpha}{\alpha} J_2 \geq 0 \tag{8}$$

$$\frac{\alpha}{1-\alpha} J_2 + \frac{1-\alpha}{\alpha} J_1 \geq 0 \tag{9}$$

This results in

$$\frac{\alpha}{1-\alpha} J_2 \leq J_1 \leq \frac{1-\alpha}{\alpha} J_2 \tag{10}$$

$$J_2 \geq 0 \tag{11}$$

and is shown graphically in figure XXXX.

In order to obtain the distribution of a single intensity of a twinned data set, integrate out J_1:

$$f(J_2) = \int_{\frac{\alpha}{1-\alpha}J_2}^{\frac{1-\alpha}{\alpha}J_2} f(J_1, J_2) dJ_1 \quad (12)$$

$$= \frac{\exp\left[-\frac{J_2}{(1-\alpha)\epsilon\sigma_p}\right] - \exp\left[-\frac{J_2}{\alpha\epsilon\sigma_p}\right]}{(1-2\alpha)\epsilon\sigma_p} \quad (13)$$

The moments of this distribution are equal to

$$\mathbb{E}[J_2] = \frac{(\epsilon\sigma_p(1-\alpha))^{1+k} - (\epsilon\sigma_p\alpha)^{1+k}}{\epsilon\sigma_p(1-2\alpha)}\Gamma[1+k] \quad (14)$$

Expression (14) can be used to obtain the familiar values of the Wilson (XXXX) or Stanley (XXXX) ratios for acentric data.

A.2. Experimental errors

If one assumes that the experimentally obtained intensity is an estimator of the 'true', error-free intensity, one can write

$$f(J|J^{\rm obs}, \sigma_{\rm obs}) = \frac{1}{\sqrt{2\pi}\sigma_{\rm obs}} \exp\left[-\frac{(J-J^{\rm obs})^2}{2\sigma_{\rm obs}}\right] \quad (15)$$

Following the approach outlined by Pannu & Read (XXXX), the distribution taking into account experimental errors can be obtained as follows:

$$f(J_1^{\rm obs}, J_2^{\rm obs}|\alpha) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(J_1, J_2|\alpha)\mathbf{I}_A(J_1, J_2, \alpha) \times f(J_1|J_1^{\rm obs}, \sigma_{\rm obs1}) f(J_2|J_2^{\rm obs}, \sigma_{\rm obs2}) dJ_1 dJ_2 \quad (16)$$

where $\mathbf{I}_A(J_1, J2)$ is an indicator function:

$$\mathbf{I}_A(J_1, J_2) = \begin{cases} 1 \text{ if } & (\frac{\alpha}{1-\alpha}J_2 \le J_1 \le \frac{1-\alpha}{\alpha}J_2) \wedge (J_2 \ge 0) \\ 0 \text{ if } & \text{otherwise} \end{cases} \quad (17)$$

One of the two integrals can be carried out analytically. The second integral can be efficiently approximated using a Gauss-Hermite or Gauss-Legendre quadrature (XXXX).

Appendix B
Dependent twin related amplitudes

B.1. Basic distributions

If due to the presence of non crystallographic symmetry twin related amplitudes can no longer be considered to be independent, the correlation between the intensities needs to be taken into account specifically. This can be done in a rather straightforward manner using the multivariate complex-normal distribution (XXXX,XXXX).

Assuming that the twin related structure factors can be described by a bivariate complex normal distribution, one can write

$$f(\mathbf{F}_1, \mathbf{F}_2) = \frac{1}{|\Sigma|} \exp\left[- \begin{pmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \end{pmatrix}^H \Sigma^{-1} \begin{pmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \end{pmatrix} \right] \tag{18}$$

With

$$\Sigma = \begin{pmatrix} \mathbb{E}[\mathbf{F}_1\mathbf{F}_1^*] & \mathbb{E}[\mathbf{F}_2\mathbf{F}_1^*] \\ \mathbb{E}[\mathbf{F}_1\mathbf{F}_2^*] & \mathbb{E}[\mathbf{F}_2\mathbf{F}_2^*] \end{pmatrix} \tag{19}$$

$$= \begin{pmatrix} \epsilon\sigma_p & \epsilon\sigma_p D_{\rm ncs} \\ \epsilon\sigma_p D_{\rm ncs} & \epsilon\sigma_p \end{pmatrix} \tag{20}$$

$D_{\rm ncs}$is multiplier similar to the Luzatti D value (XXXX), with the difference that it is measure of similarity between the model and its NCS related counterpart. Of course $0 \geq D_{\rm ncs} \geq 1$. Transforming to polar coordinates, integrating out phases and changing variables from amplitudes to intensities, one obtains

$$f(I_1, I_2) = \frac{1}{(\epsilon\sigma_p)^2(1 - D_{\rm ncs}^2)} \exp\left[- \frac{I_1 + I_2}{\epsilon\sigma_p(1 - D_{\rm ncs}^2)} \right] \times I_0\left[\frac{2D_{\rm ncs}\sqrt{I_1 I_2}}{\epsilon\sigma_p(1 - D_{\rm ncs}^2)} \right] \tag{21}$$

Note that

$$\mathbb{E}[I_.] = \epsilon\sigma_p \tag{22}$$

$$\mathbb{E}[I_.^2] = 2(\epsilon\sigma_p)^2 \tag{23}$$

$$\mathbb{E}[I_1 I_2] = (1 + D_{\rm ncs}^2)(\epsilon\sigma_p)^2 \tag{24}$$

and thus

$$\begin{aligned} {\rm CC}[I_1, I_2] &= \frac{\mathbb{E}[I_1 I_2] - \mathbb{E}[I_1]\mathbb{E}[I_2]}{\sqrt{(\mathbb{E}[I_1^2] - \mathbb{E}[I_1]^2)(\mathbb{E}[I_2^2] - \mathbb{E}[I_2]^2)}} \\ &= D_{\rm ncs}^2 \end{aligned} \tag{25}$$

$$\begin{aligned} {\rm R}^{\rm sq} &= \frac{\mathbb{E}[(I_1 - I_2)^2]}{\mathbb{E}[(I_1 + I_2)^2]} \\ &= \frac{1 - D_{\rm ncs}^2}{3 + D_{\rm ncs}^2} a \end{aligned} \tag{26}$$

B.2. Twinning

Given expression (21), twinning can be introduced in exactly the same manner as done in section(XXXX). The resulting distribution is

$$\begin{aligned} f(I_1, I_2|\alpha, D_{\rm ncs}) &= \frac{1}{(\epsilon\sigma_p)^2(1 - D_{\rm ncs}^2)(1 - 2\alpha)} \times \\ &\quad \exp\left[-\frac{J_1 + J_2}{\epsilon\sigma_p(1 - D_{\rm ncs}^2)}\right] I_0\left[X\right] \end{aligned} \tag{27}$$

with

$$X = \frac{2D_{\rm ncs}\sqrt{((1-\alpha)J_1 - \alpha J_2)((1-\alpha)J_2 - \alpha J_1)}}{\epsilon\sigma_p(1 - D_{\rm ncs}^2)(1 - 2\alpha)} \tag{28}$$

The domain on which this distribution is defined is as in expression (11)

In order to asses the effect of twinning on $R^{\rm sq}$, expression (26) note that

$$J_1 - J_2 = (1 - 2\alpha)(I_1 - I_2) \tag{29}$$

$$J_1 + J_2 = (I_1 + I_2) \tag{30}$$

It is then easy to see that

$$\begin{aligned} \mathrm{R}_{\mathrm{twin}}^{\mathrm{sq}} &= \frac{\mathbb{E}[(J_1 - J_2)^2]}{\mathbb{E}[(J_1 + J_2)^2]} \\ &= (1-2\alpha)^2 \frac{1 - D_{\mathrm{ncs}}^2}{3 + D_{\mathrm{ncs}}^2} \end{aligned} \tag{31}$$

B.3. Introducing experimental errors

Introduction of experimental errors is done as for the independent case:

$$\begin{aligned} f(J_1^{\mathrm{obs}}, J_2^{\mathrm{obs}} | \alpha, D_{\mathrm{ncs}}) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(J_1, J_2 | \alpha, D_{\mathrm{ncs}}) \times \\ & \quad \mathbf{I}_A(J_1, J_2) f(J_1 | J_1^{\mathrm{obs}}, \sigma_{\mathrm{obs1}}) \times \\ & \quad f(J_2 | J_2^{\mathrm{obs}}, \sigma_{\mathrm{obs2}}) dJ_1 dJ_2 \end{aligned} \tag{32}$$

Unfortunately, no analytical solution could be found for these integrals. A numerical approach using a combination of Gauss-Hermite of Gauss-Legendre quadratures proved to be a feasable alternative.

References

Author, A. & Author, B. (1984). *Journal* **Vol**, first page–last page.

Fig. 1. The distribution of normalized structure factor amplitudes for various twin fractions.

Fig. 2. A graphical depiction of the domain on which the distribution of the twinned intensities (J_1, J_2) is defined. The arrows indicate the domain where $\frac{\alpha}{1-\alpha} J_2 le J_1 \leq \frac{1-\alpha}{\alpha} J_2$

Fig. 3. The dependce of $R^{\rm sq}_{\rm twin}$ on the twin fraction α and $D_{\rm ncs}$. The green broken line indicate the isoline for which $R^{\rm sq}_{\rm twin} = 1\%$

Synopsis

Likelihood based techniques for determining the twin fraction are described.